Building a Central Knowledge Graph for Multiple Humanities Research Projects

Steffen Hennicke[1], Kim Pham[1], Robert Casties[1], and Esther Chen[1]

[1]Max Planck Institute for the History of Science

1 Management of Digital Research Data

Research projects in the Humanities use and produce many different kinds of research data that need to be preserved after the project has ended. While libraries and archives as well as museums have traditionally taken care of analogue research data such as printed monographs, archival documents, and physical objects, there remain challenges in dealing with the rapidly increasing amounts of digital research data.

In digital humanities projects, digital artefacts — such as digitised sources, databases or datasets, project websites, (analytic) algorithms, or (interactive) visualisations —, often constitute the primary foundation of research results or even are their primary outcome. One of the main challenges is that digital research data and programming logic as well as presentation have often been entangled.

As a result, the preservation of digital research data for long-term access and subsequent re-use often becomes problematic, because of the lack of technical standards and its dependence on outdated software. In these situations valuable data becomes unreadable and are unable to be reused. Digital research data, however, must be designed to be sustainable in order "to keep them alive" (Kilchenmann et al., 2019).

To address this challenge, a team of librarians and IT researchers at the Max Planck Institute for the History of Science (MPIWG) began to develop DRIH, the digital research infrastructure for the humanities – that is meant to secure research data generated in the many projects of the institute.

2 Graceful Degradation

Our thinking is based on the conceptualisation of research data as digital artefacts that need to be either stored, hosted, or run.[1] Digital Artefacts that need to be run or hosted — such as dynamic websites, interactive visualisation, or code — have specific software dependencies that diminish their preservation life-time, that is the time they can be reasonably kept running and working as originally intended due to increasingly high maintenance of software dependencies.

Digital artefacts that can be stored are any individual or collection of digital files. If properly stored and archived in a non-proprietary file format, and with their semantics

[1] For the following also cf. Kräutli (2021).

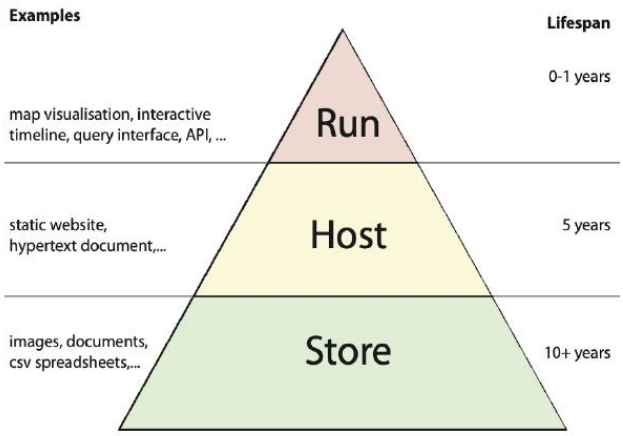

Figure 1: Lifespans of Different Types of Digital Artefacts.

well modelled and documented, then these research data remain accessible over the long-term and even allow to be integrated with data from different projects and thereby opening up new possibilities and synergies.

Therefore, our general approach for dealing with digital projects and their research data at the MPIWG "is to produce [digital] artefacts that can be gracefully degraded to the lower and more sustainable parts of the store/host/run pyramid" (Kräutli, 2021), as shown in Figure 1, and to store and archive those data in a way, that makes them re-usable in the long-run.

3 DRIH

To realise this core principle of preserving research data, that is, to separate data from software and presentation, and to manage the data outputs of projects' end-of-life, we are building a digital research infrastructure for the humanities (DRIH).

The infrastructure is comprised of four primary systems with four primary functions (see Figure 2): (1) digitisation and ingest of digital objects and bibliographic information (Goobi[2]), (2) storage and preservation of digital objects (CDStar[3]), (3) digital repository for digital object management (DRIH repository with Fedora), (4) building a central Knowledge Graph. These systems are integrated by a suite of microservices, plugins, and scripts that also help to automate, deploy, and operate the infrastructure.

Two main components support the development of a central Knowledge Graph. The digital repository is the first component, composed of a number of plugins such as Solr for indexing content, using the IIIF standard to manage media, Fedora to organize metadata, epic for our identifier service, X3ML for mapping metadata to linked data, and a custom administrative interface. The second component is the central knowledge graph system, which is built using Blazegraph and the Metaphactory research platform that preserves research data in a flexible semantic data format called CIDOC CRM.

When objects are digitized and added to the DRIH system, its metadata is first ingested as METS/MODS. The object and metadata are stored and transformed using

[2] https://www.intranda.com/en/digiverso/goobi/goobi-overview/
[3] https://cdstar.gwdg.de/

the X3ML mapping engine, which uses our custom mappings to transform it to RDF using the CIDOC CRM standard. These triples can be viewed and explored along with the digital objects in the knowledge graph frontend.

The Knowledge Graph addresses the bottom of the pyramid from Figure 1 and archives research data as CIDOC CRM and Linked Data. In the following, we want to first discuss, why we use CIDOC CRM and Linked Data, and then report on our initial use cases on mapping and modelling project data and their context for the Knowledge Graph.

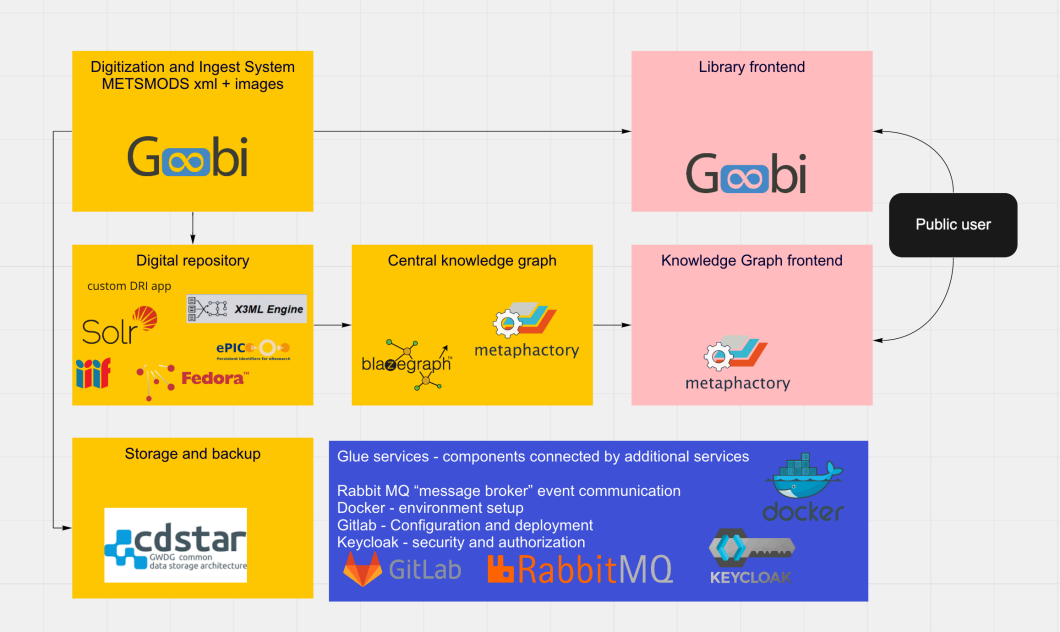

Figure 2: DRIH architecture and technologies used to build the infrastructure.

4 Central Knowledge Graph

The purpose of the Knowledge Graph is (1) to store and archive project's research data as sustainable and reusable datasets, (2) to provide an interface for querying (a) the contents of all project's research data across datasets' boundaries and (b) to identify any digital artefacts that constitute a relevant research output of a research project, understand its provenance, and find the location where it is being hosted, run, or stored. That means that while the Knowledge Graph on the one hand stores and archives research data as re-expressions in CIDOC CRM, it functions, at the same time, as a registry for any other digital artefact that has been produced by a research project. The Knowledge Graph, therefore, constitutes a key component in the Research Data Management strategy of the institute.

4.1 CIDOC CRM and Linked Data

We have chosen CIDOC CRM as the common semantic target model. The CIDOC CRM is a formal ontology that defines "the underlying semantics (…) used in the documentation of cultural heritage and scientific activities" (Bekiari et al., 2021, 9). Its scope covers the "factual knowledge about the past at a human scale" (Bekiari et al., 2021, 10). CIDOC CRM is widely used by cultural heritage institutions, and

provides a sustainable foundation for data interoperability (Doerr and Iorizzo, 2008), which is a key challenge for ingesting data from different humanities projects.

The CIDOC CRM allows to re-express the complexity and intricacies, that are contained in the various project data that we ingest from the institute's projects. For this purpose, the CIDOC CRM provides a high-level semantic, integrative layer, but allows, at the same time, to specialise its classes and properties, so that project data may define specific meanings while still retaining interoperability at the high-level of the ontology.

CIDOC CRM as such is agnostic to a specific data format representation. However, the most prominent implementation is based on Linked Data technologies (Heath and Bizer, 2011) that makes CIDOC CRM immediately machine-readable and easily connectable to other data sources on the Semantic Web. Together, CIDOC CRM and Linked Data provide the means to create semantically and structurally integrated and interoperable research data, that are easily linked, sustainable, and reusable. We have found that CIDOC CRM and its extensions appear to be well suited to serve as a common target model, that is to accommodate the diverse data outputs that are produced at the MPIWG and to re-express their intrinsic complexity. In the following we discuss some of our experiences so far with mapping and re-expression of project data with CIDOC CRM and Linked Data.

4.2 Mapping and Modelling of Project Data

As first use cases we have worked on models of complex network data from two humanities research projects, The Sphere[4], a project that traces knowledge development through the early modern publication history of Johannes de Sacrobosco's De Sphaera and the ISMI[5] project, that collects information on all Islamic manuscripts in the exact sciences from the 9th to the 19th century, as well as bibliographic resource data from the MPIWG library.

These three projects represent three different use-cases: A project with data born in CIDOC CRM, a project where the full workflow was migrated to CIDOC CRM, and a project with a different workflow where only the output is converted to CIDOC CRM.

The Sphere project data model was created using CIDOC CRM with some extensions from FRBRoo (Bekiari et al., 2017) and CRMdig[6] (Doerr et al., 2016) from the beginning and all data was entered using a frontend based on Metaphacts Open Source Platform[7] configured using this data model.

The ISMI project is an important and long-term institutional project that started out with a custom graph-like data model and a custom database in 2005 and was migrated to CIDOC CRM towards the end of the projects' time at the MPIWG. The migration process included the creation of a CIDOC CRM data model and a mapping of the projects' custom data format and model to CIDOC CRM using the X3ML[8] tool and the creation of a new Metaphacts Platform-based frontend for data entry. Special caution was taken to ensure the successful migration of the complex data graph by creating an additional mapping from CIDOC CRM back into the projects' custom format to compare the output of the migration round-trip with the original data.

---

4 https://sphaera.mpiwg-berlin.mpg.de/ [2021-09-13]

5 https://ismi.mpiwg-berlin.mpg.de/ [2021-09-13]

6 https://cidoc-crm.org/frbroo/, https://cidoc-crm.org/crmdig/

7 https://bitbucket.org/metaphacts/metaphacts-community/

8 https://github.com/isl/x3ml

The bibliographic data of the digitised rare books collection of the MPIWG library is managed using the Goobi workflow tool in METS/MODS format and it is transformed to CIDOC CRM using the X3ML tool. This dataset is a useful resource on its own but it also contains books that have been used in other projects like ISMI and Sphere, and we want to use it to demonstrate how datasets can become interconnected.

Our initial experience shows that CIDOC CRM with some extensions allows us to model the salient aspects of these data sets created as part of research in the history of science quite well but the process of data modelling requires a serious amount of work by modelling and domain specialists both initially and over time. Another challenge is presented when a project requires the ongoing creation and modification of the graph in a triplestore through a simple frontend as opposed to more established frontends built on relational databases. The LDP-based approach of Metaphacts Platform forms and named graphs deliver a viable frontend solution but requires additional concern in data modelling and management.

4.3 Project-Level Description

When employing a Knowledge Graph as a tool for research data management one important challenge is to address not only querying for the contents of project datasets but also to provide the necessary information that allows users to understand and identify the projects' context and the provenance of its digital artefacts. For this purpose, we are currently modelling a superordinate descriptive layer on top of the specific project datasets in the Knowledge Graph.

We have formulated initial requirements to describe the projects and actors, as well as the datasets, software and services project's provide or maintain. We also model our own hosting and mapping services provided as the Research Data Management team for the Knowledge Graph. In particular, we keep track of where mapping instructions and source data have been archived that have been used to generate the datasets in the Knowledge Graph. The goal is to enable the user to assess and trust the information and data that can be found in and via the Knowledge Graph.

For this purpose we have started to evaluate the Parthenos (Bruseker et al., 2017) ontology that supports a "sustainable cross-disciplinary semantic graph of available resources" and that ensures "a provenance of knowledge with regards to their epistemic status and origin" (Bruseker et al., 2017, 1). The ontology is an extension to the CIDOC CRM and is therefore immediately compatible with our CIDOC CRM model in the Knowledge Graph. The entities and relations described by the Parthenos ontology together with the CRMdig have allowed us to represent most of our requirements so far.

5 Conclusion

Our work on the DRIH infrastructure and the Knowledge Graph is on-going. Our experiences with the application and implementation of a Knowledge Graph and CIDOC CRM have been positive, although, there is a steep learning curve when it comes to semantic modelling and mapping of data to the CIDOC CRM. There remain several challenges to address, such as the implementation of the Fundamental Categories and Relationships (Tzompanaki and Doerr, 2012) that would allow to target specific information needs of user groups, named graph management, or how to deal with named entity data that have been created in projects. This kind of data in particular is

highly valuable for future research projects, since the data that has been created on, for example, persons or institutions, are specific to the general research domain of the institute, and constitute trustworthy contextual information that is typically not found in external authority data.

References

Bekiari, C., G. Bruseker, M. Doerr, C.-E. Ore, S. Stead, and A. Velios (2021). Volume A: Definition of the CIDOC Conceptual Reference Model. Produced by the CIDOC CRM. Version 7.1.1, April 2021. Produced by the ICOM/CIDOC Documentation Standards Group, continued by the CIDOC CRM Special Interest Group. Technical report.

Bekiari, C., M. Doerr, P. Le Boeuf, and P. Riva (2017). FRBR Object-Oriented Definition and Mapping from FRBRer, FRAD and FRSAD (Version 3.0). Technical report, International Working Group on FRBR and CIDOC CRM Harmonisation.

Bruseker, G., M. Doerr, and M. Theodoridou (2017). Report on the Common Semantic Framework.

Doerr, M. and D. Iorizzo (2008). The Dream of a Global Knowledge Network: A New Approach. 1(1), 1–23.

Doerr, M., M. Theodoridou, and S. Stead (2016). Definition of the CRMdig. An Extension of CIDOC-CRM to Support Provenance Metadata (3.2.1). Technical report, FORTH.

Heath, T. and C. Bizer (2011). Linked Data: Evolving the Web into a Global Data Space, Volume 1 of Synthesis lectures on the semantic web theory and technology. San Rafael: Morgan & Claypool.

Kilchenmann, A., F. Laurens, and L. Rosenthaler (2019). Digitizing, Archiving... and then? Ideas about the Usability of a Digital Archive. Archiving Conference 2019(1), 146–150.

Kräutli, F. (2021). Digital Research Afterlife: Graceful Degradation in Digital Humanities Projects.

Tzompanaki, K. and M. Doerr (2012). Fundamental Categories and Relationships for Intuitive Querying CIDOC-CRM Based Repositories. Technical Report TR-429, ICS-FORTH, Heraklion. Issue: TR-429.